Desempeño de un modelo de lenguaje amplio de inteligencia artificial en un examen habilitante en Anestesiología de la Corporación Nacional de Certificación de Especialidades Médicas (CONACEM)

https://doi.org/10.25237/congreso2023-2

Fernando R. Altermatt Couratier[1], Hector J. Lacassie Quiroga[1], Andrés Neyem[2]

[1]División de Anestesiología, Facultad de Medicina, Pontificia Universidad Católica de Chile.

[2]Departamento de Ciencia de la Computación, Facultad de Ingeniería, , Facultad de Medicina, Pontificia Universidad Católica de Chile.

Introducción

ChatGPT (OpenAI; San Francisco, CA) es un tipo de inteligencia artificial (IA) en la forma de un modelo de procesamiento de lenguaje natural amplio, al que se puede acceder libremente.

Recientemente la última versión (ChatGPT-4) ha demostrado la capacidad de aprobar el examen de especialidad de anestesiología estadounidense (American Board of Anesthesiology, con 78%) y una muestra del examen de especialidad del Royal College of Anaesthesiologists, Reino Unido (63%), resultados que hasta ahora no se habían logrado con versiones previas.

En Chile, para certificar la calidad de especialista de profesionales egresados de universidades no acreditadas en el país, se requiere de un examen de especialidad. Para su aprobación, se debe lograr un porcentaje de respuestas correctas de al menos 65%.

Objetivos

Nuestro objetivo es evaluar si las versiones de ChatGPT-3,5 y 4 son capaces de aprobar un examen de especialidad en anestesiología chileno.

Materiales y Métodos

El examen de especialidad de anestesiología de CONACEM es un examen habilitante, de múltiple elección, que incluye preguntas de todas las áreas de la Anestesiología. Se realiza de forma presencial y los participantes disponen de tres horas para realizarlo. Utilizamos el examen de noviembre de 2018. Las preguntas fueron contestadas por ChatGPT versión 3,5 y 4. Luego, las preguntas fueron traducidas al idioma inglés, utilizando el mismo programa y se repitió el proceso para determinar si el idioma podía ser una limitante en términos de resultados.

Resultados

El examen de CONACEM de noviembre 2018 fue aprobado en esa ocasión por 49 de un total de 101 postulantes (49%). El promedio de respuestas correctas fue de 44 (intervalo de confianza 95%: 42,2-45,7).

En junio de 2023 se ejecutaron los experimentos. En todas las instancias, ChatGPT-3,5 no logró superar el umbral de 65% de respuestas correctas, tanto en español como en inglés. Para el caso de ChatGPT-4, logró aprobar el examen en español e inglés, donde aparentemente mejoró su desempeño, sin embargo, no fue estadísticamente significativo (tabla).

Conclusiones y/o implicaciones

Esta es la primera experiencia evaluando el rendimiento de una inteligencia artificial en la forma de un modelo de procesamiento de lenguaje natural amplio, con un examen habilitante de especialidad médica en idioma español. Nuestro principal hallazgo es que ChatGPT-4 logró superar el umbral de aprobación del examen de certificación de especialidad en anestesiología de 2018 en Chile, similar a la cohorte que rindió el examen.

Experiencias con otros exámenes de licenciamiento han sido exitosos, lo que se repite en nuestra experiencia, aunque sólo con la versión más actual (y pagada) de ChatGPT. La barrera idiomática fue descartada como factor determinante.

Estamos en los albores de la inteligencia artificial en su forma de aprendizaje de máquinas, sin embargo, hemos visto su rápida evolución, que emula e incluso supera el desempeño humano. Futuros estudios deben indagar en la capacidad de razonamiento de estos modelo de procesamiento de lenguaje natural amplio, con el objeto de precisar cuáles son los modelos analíticos con los que estas plataformas abordan la resolución de problemas clínicos.

Referencias

1.Angel, M. C., Rinehart, J. B., Canneson, M. P. & Baldi, P. Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the ABA Exam. medRxiv 2023.05.10.23289805 (2023) doi:10.1101/2023.05.10.23289805.2.Aldridge, M. J. & Penders, R. Artificial intelligence and anaesthesia examinations: exploring ChatGPT as a prelude to the future. Br. J. Anaesth. (2023) doi:10.1016/j.bja.2023.04.033.

Gráficos, Tablas e Imágenes

ChatGPT-3,5 (español)	nov-18
CORRECTAS	28
ERRÓNEAS	42
NOTA	3

ChatGPT-3,5 (inglés)	nov-18
CORRECTAS	33
ERRÓNEAS	37
NOTA	3,2

ChatGPT-4 (español)	nov-18
CORRECTAS	47
ERRÓNEAS	23
NOTA	4,2

ChatGPT-4 (inglés)	nov-18
CORRECTAS	53
ERRÓNEAS	17
NOTA	4,9

Tabla. Comparación de resultados de ChatGPT-3,5 y 4 para el examen de noviembre 2018 de Conacem en español y en inglés. Nota 4 (aprobación) requiere 65% de correctas.